DOCUMENT RESUME

ED 346 160 TM 018 521

AUTHOR Hennings, Sara S.; Hughes, Kay E.
TITLE Building a Performance-Based Assessment System To Diagnose Strengths and Weaknesses in Reading Achievement.
PUB DATE Apr 92
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Diagnostic Tests; *Elementary School Students; Elementary Secondary Education; *Evaluation Methods; Field Studies; Language Arts; Oral Reading; Reading Achievement; Reading Comprehension; *Reading Diagnosis; Reading Tests; *Secondary School Students; Semantics; Silent Reading; Spelling; Student Evaluation; *Test Construction; Test Validity; Word Recognition
IDENTIFIERS *Diagnostic Assess Reading Trial Teach Strategies; *Performance Based Evaluation

ABSTRACT

This paper provides a brief description of the development of the Diagnostic Assessments of Reading with Trial Teaching Strategies (DARTTS) program by F. G. Roswell and J. S. Chall. It also describes the editorial and statistical procedures that were used to validate the program for determining students' strengths and weaknesses in important areas of reading and language. The DARTTS program is a unique performance-based assessment that serves as a testing/teaching diagnostic tool. The Diagnostic Assessments of Reading are short formal tests that assess strengths and weaknesses in six areas of reading and language: (1) word recognition; (2) word analysis; (3) oral reading; (4) silent reading comprehension; (5) spelling; and (6) word meaning. The informal assessment lessons of the Trial Teaching Strategies are organized by developmental levels from beginning through advanced. The tests were developed and given an initial item try-out in 1989 with 1,664 students in grades 2 through 8. A 1990-91 validation study involved 1,216 students in grades 1 through 12. The process of item development is illustrated by referring to a specific test section. The procedures that validated the DARTTS program may have applications in developing other similar instruments. There are 12 tables of study data and an 10-item list of references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED346160

Building a Performance-Based Assessment System to Diagnose Strengths and Weaknesses in Reading Achievement

by
Sara S. Hennings
Kay E. Hughes

The Riverside Publishing Company
8420 W. Bryn Mawr Avenue
Chicago, IL 60631

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

SARA S. HENNINGS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

D·A·R·T·T·S™
Diagnostic Assessments of Reading with Trial Teaching Strategies
Florence G. Roswell and Jeanne S. Chall

American Educational Research Association
Annual Meeting
San Francisco, CA
April 1992

TM018521

BEST COPY AVAILABLE

Building a Performance-Based Assessment System to Diagnose Strengths and Weaknesses in Reading Achievement

This paper provides a brief description of the development of the *Diagnostic Assessments of Reading with Trial Teaching Strategies* (*DARTTS*™) program, by Florence G. Roswell and Jeanne S. Chall. It then describes in detail the editorial and statistical procedures that were employed to validate the use of the program for the purpose of determining students' strengths and weaknesses in important areas of reading and language. Parts of the text and data have been taken from the *DARTTS Technical Manual* (Roswell & Chall, in press).

Part I: Description of the *DARTTS* Program

The *Diagnostic Assessments of Reading with Trial Teaching Strategies* (*DARTTS*) program is a unique performance-based assessment which serves as a testing/teaching diagnostic tool. It is designed for use by classroom teachers, reading specialists, special education teachers, and other professionals who have interest in helping students read better. The *Diagnostic Assessments of Reading* (*DAR*) are short, formal tests that assess students' relative strengths and weaknesses in six areas of reading and language. The *Trial Teaching Strategies* (*TTS*) are brief informal assessments in the form of lessons based on the results of the *DAR* tests. During these lessons both the teacher and the student learn about effective procedures and materials that can be used to overcome the student's reading difficulties. Such knowledge serves as a powerful incentive in engaging the student's cooperation and interest in his or her newly planned instructional program. The emphasis in the *DARTTS* program is on the individual student, on aiding the student to read better so that he or she will enjoy reading and be able to use constructively his or her potential reading abilities.

The formal tests of the *Diagnostic Assessments of Reading* (*DAR*) are as follows:

- Word Recognition—a test of reading vocabulary that also serves to indicate the appropriate starting point for the other *DAR* tests (except Word Meaning)
- Word Analysis—a test of letter/sound correspondences from single consonants through a range of phonic knowledge that is administered to students who score at Level 3 or below on the Word Recognition test
- Oral Reading—a test that suggests the appropriate level for reading connected text and reveals types of students' errors that indicate instructional needs in word analysis. It further reveals students' fluency, expression, and ability to use context intelligently.
- Silent Reading Comprehension—a test of reading comprehension that reveals how well students can extract information from text independently through both objective questions and an oral account. This broad-based approach to silent reading comprehension provides the teacher with information about the student's depth of understanding of the text and his or her language organization.
- Spelling—an orally administered test with written responses that reveals various types of spelling difficulties and also provides the teacher with an initial printing or handwriting sample from the student

- Word Meaning—an orally administered test of meaning vocabulary that serves as a measure of language development and can suggest students' learning potential, especially for younger readers

The informal assessment lessons of the *Trial Teaching Strategies* (*TTS*) are organized by developmental levels as follows:

- For Beginning Readers (*DAR* Levels 1-1 and 1-2), the *TTS* concentrate on word recognition and analysis and on oral reading.

- For Primary Readers (*DAR* Levels 2 and 3), the *TTS*, in addition to word recognition and analysis and oral reading, include a spelling lesson and a lesson in silent reading comprehension and written expression (Level 3 only).

- For Intermediate Readers (*DAR* Levels 4, 5, 6, 7, and 8), the *TTS* continue with oral reading (Levels 4, 5, and 6 only) and spelling, and concentrate on silent reading comprehension and written expression, including textbook analysis and study skills.

- For Advanced Readers (*DAR* Levels 9/10 and 11/12), the *TTS* emphasize silent reading comprehension and written expression, textbook analysis and study skills, and spelling.

Informal lessons on vocabulary development are provided for all levels.

Table 1 shows the multilevel organization of the *DAR*. Table 2 provides an overview of the lessons of the TTS.

Table 1

Organization of the *Diagnostic Assessments of Reading* (*DAR*)

DAR Test	*DAR* Level										
	1-1	1-2	2	3	4	5	6	7	8	9/10	11/12
Word Recognition	x	x	x	x	x	x	x	x	x	x	x
Word Analysis[1]	x	x	x	x							
Oral Reading	x	x	x	x	x	x	x	x	x	x	x
Silent Reading Comprehension[2]				x	x	x	x	x	x	x	x
Spelling	x	x	x	x	x	x	x	x	x	x	x
Word Meaning[3]		x		x	x	x	x	x	x	x	x

[1] The Word Analysis test is administered only to students who score at Level 3 or below on the Word Recognition test.

[2] The Silent Reading Comprehension test is administered only to students who score at Level 3 or above on the Word Recognition test.

[3] There is only one level of the Word Meaning test at the lowest level.

Table 2

Overview of the *Trial Teaching Strategies* (*TTS*)

DAR Level	*TTS* Level	*TTS* Lesson
1-1 1-2	Beginning Readers	Word Recognition and Analysis Oral Reading
2 3	Primary Readers	Word Recognition and Analysis Oral Reading Silent Reading Comprehension and Written Expression Spelling
4 5 6 7 8	Intermediate Readers	Oral Reading Silent Reading Comprehension and Written Expression Spelling
9/10 11/12	Advanced Readers	Silent Reading Comprehension and Written Expression Spelling

The teacher is instructed to begin the *DAR* Word Recognition test at the student's current functioning level in reading, regardless of grade placement in school. If available, the results of a standardized test of reading can be used. However, an estimate based on the teacher's general knowledge of the student's current functioning in reading can also be used. The *DARTTS* program refines this information for the student and teacher, and so an approximation of the student's reading level is sufficient to begin testing. If a student fails to achieve mastery on the first test level attempted, the teacher is instructed to present the next lower level of the test until the student does achieve mastery on one level. For each test in the *DAR*, the last level a student *masters* that is followed by a level that the student *does not master* is the student's "highest mastery level" for that test. The simple focus of the *DAR* is to discover the student's highest mastery level for each of the areas of reading and language that are tested.

For all of the *DAR* tests, mastery is defined as correct responses on approximately 70% of the items for that test level. Specific mastery criteria based on this standard are provided in the testing materials, both in the *DAR Teacher's Manual* and in the *DAR Directions for Administration with*

Response Record, which is the teacher's complete record of the student's responses on the *DAR*. For example, mastery for the Word Recognition test is reading correctly seven out of the ten words presented at any given level.

The student's highest level of mastery for the Word Recognition test indicates the appropriate starting level for the *DAR* Word Analysis, Oral Reading, Silent Reading Comprehension, and Spelling tests. For Word Meaning, because the student is not required to do any reading or writing, the teacher is instructed to begin at the student's current grade placement. For all tests, the administration continues until the student reaches a highest mastery level—that is, until the student masters one level and then fails to achieve mastery on the next higher level.

Based on the results of the *DAR*, the teacher creates a profile of the student's strengths and weaknesses in the six areas of reading and language that are tested. Then the teacher and student together try the appropriate lessons of the *Trial Teaching Strategies* (*TTS*). For the *DAR*, the teacher is instructed to follow the testing procedures established during the national Validation Study. For the *TTS*, on the other hand, the teacher is told to provide the student with unrestricted support of his or her efforts to learn.

Based on the *DAR* score profile and the teacher's qualitative evaluation of the student's reaction to and success with the lessons of the *TTS*, a program of remedial instruction is planned for the individual student. This planning can also include the placement of students who have similar instructional needs into small groups for the most efficient use of the teacher's time.

Program Authors

Florence G. Roswell is Professor Emerita, City College, City University of New York. While at City College, Professor Roswell engaged in teacher education and served in a consultant capacity to schools regarding their reading programs. The director of the City College Reading Center for 22 years, she is the author of *Reading Disability: A Human Approach to Evaluation and Treatment of Reading and Writing Difficulties,* now in its fourth edition. Professor Roswell has had wide experience as a clinical psychologist specializing in diagnosis and treatment of reading disability. She is currently a consultant to reading professionals.

Jeanne S. Chall has been actively involved in reading research, teacher education, and clinical practice. She was professor and director of the Harvard University graduate program in reading and language and the director of the Harvard Reading Laboratory for 25 years. She is author of many books on reading, including *Stages of Reading Development* and *The Reading Crisis*. Professor Chall is regularly called upon to advise various governmental and educational groups and school systems regarding issues related to reading and reading instruction.

Manuscript Development

In both their clinical work and research, Professors Roswell and Chall have always depended on reliable tests for diagnostic information. Many years ago when they first worked together, because no single published instrument could meet their needs, they used the best of the available tests. However, using many different tests took a great deal of time and still did not provide all of the information they needed. They determined that they needed to create a single, shorter test which would provide them with all of the diagnostic information they required

in the least amount of administration time. One concern, of course, was to limit testing time as much as possible. However, they were also concerned about preserving the self-esteem of their students. A strong theme that runs throughout the *DARTTS* program materials is the importance of treating students with respect and dignity and facilitating a success experience for them as soon as possible in order to enlist their cooperation with the remediation process. Because Roswell and Chall could find no commercially available, short, reliable test that would yield the information they required, they began the development of their own test. After considerable research and further refinements, the *DARTTS* program evolved.

Initial Prepublication Research—The 1989 National Item Tryout

Even though an early version of the *DARTTS* program had been researched by the authors within the sphere of their own clinical and research activities (Chall, Jacobs, & Baldwin, 1990), the publisher decided to conduct a national item tryout on the formal assessment component—the *DAR*—to compile a pool of researched items from which final test items could be selected. For this national item tryout, two complete forms of the test were created. These two forms, Form X and Form Y, were tried out in fall 1989 with 1664 students in grades 1–8.

Table 3 shows the numbers of students by grade that were included in the Tryout Study.

Table 3

Numbers of Students in the *DAR* Tryout Study

Grade	Form X	Form Y
1	169	123
2	148	170
3	134	122
4	110	106
5	108	108
6	98	77
7	40	61
8	50	40
Total	857	807

Teachers were instructed to select students for the *DAR* tryout who were reading approximately *on grade level*. Students who could read on grade level would be needed for the purpose of establishing empirically the grade level of the test items. The final *DAR* has a multilevel format so that teachers can easily move forward and backward through the test levels. However, because tryout participants would be reading on grade level, to make the data easier to analyze, special grade-level *DAR* tryout test booklets were created that had a fixed number of items. The grade 1 test booklet included the two test levels for grade 1 (Level 1-1 and Level 1-2) and also four levels of the *DAR* Word Meaning test. The test booklets for grades 2–4 included the on-grade tests plus those for two levels below and one above grade level. The test booklets for grades 5–8 included the tests for one test level below and two above in addition to the on-grade tests. This arrangement of levels ensured that tryout data for any one test level (except Level 11/12) were obtained for a minimum of two grade levels. These data were used to select the reading passages and test items that possessed the best statistical characteristics for the final test form.

Teachers were asked to administer the entire tryout test booklet to participating students. So that students would not become excessively frustrated by items that were much beyond their current level of functioning, instructions were to stop a test immediately and go on to the next test in the booklet when students missed a certain number of items in a row. (This number varied by test.) In these cases, the data analysis made the assumption that the student would have answered incorrectly all of the items from the point where the teacher stopped.

Along with the *DAR* tryout tests, teachers in grades 2–8 also administered the Vocabulary test of the on-grade level of the *Gates-MacGinitie Reading Tests* (*GMRT*) to participating students. The Vocabulary portion of the *GMRT* was chosen as the tryout anchor over the whole test primarily as a time-saving convenience for research participants. The Vocabulary test takes just 20 minutes to administer (as opposed to 55 minutes for the entire test), and it correlates strongly with the Comprehension test (MacGinitie & MacGinitie, 1989).[1] Correlations between the results of the *DAR* tryout tests and the students' scores on the *GMRT* Vocabulary test are reported in Table 4.

In grade 1, the Letters and Letter-Sound Correspondences subtest of Level PRE of the *Gates-MacGinitie Reading Tests* was administered as the research anchor test. The format of Level PRE, which tests students' knowledge of pre-reading concepts, is very different from the format of the other *GMRT* levels. The correlations reflect this format difference. Level PRE is not a reading test as such, and so the correlations with the reading tasks of the *DAR* at grade 1, as reported in Table 4, are not as high as they are for the other grade levels.

[1] The average correlation in grades 2–8 from the fall standardization of the *GMRT* is 0.79.

The correlations between the *GMRT* Vocabulary test and the *DAR* Word Meaning test are not as strong as those for the *GMRT* and the *DAR* measures of reading—Word Recognition, Word Analysis (grades 2 and 3), and Silent Reading Comprehension. The *GMRT* Vocabulary test requires reading of the student; the *DAR* Word Meaning test does not. Indeed, this characteristic of the *DAR* is a particular strength, since the Word Meaning test helps determine the potential reading abilities of the students. Spelling correlations are mixed, probably because this test is primarily a test of language, even though some reading is required.

At all levels, the correlations between the *GMRT* and the *DAR* Oral Reading tryout test show negative values. These values are a function of the scoring of the Oral Reading test, which is scored as the number of errors (reading miscues), as opposed to the number correct. Some of the apparent inconsistencies in the correlations can be explained by the characteristics of the student sample. For example, at grade 7 the students in the Form Y tryout sample were slightly more able than those in the Form X sample. (See Table 5.) Other variations are not so easy to explain.

Table 5 contains the *GMRT* Vocabulary test means and standard deviations for the *DAR* tryout and the *GMRT* national standardization samples. It also gives the same information for the DARTTS Validation Study, which is described on pages 9–13. The numbers of students in all *DARTTS* samples are also provided. This table illustrates that the students in the Tryout Study were generally reading on grade level based on their *GMRT* Vocabulary test Extended Scale Scores (ESSs). The Validation Study participants, on the other hand, were generally reading below grade level. See Table 7 on page 10 for further information on the characteristics of the Validation Study sample.

Table 4

Correlations between the Vocabulary Test of the *Gates-MacGinitie Reading Tests* and the *DAR* Tests by Form and Grade
DARTTS Tryout Study

	Form X							Form Y						
Grade	WR	WA	OR*	SRC	SP	WM	N	Grade	WR	WA	OR*	SRC	SP	WM	N
1**	0.20	0.29	-0.07		0.28	0.03	141	1**	0.08	0.45	-0.02		0.12	0.14	87
2	0.80	0.73	-0.40		0.69	0.37	135	2	0.78	0.72	-0.55		0.66	0.21	15?
3	0.66	0.63	-0.55	0.56	0.49	0.41	115	3	0.63	0.55	-0.46	0.37	0.42	0.34	104
4	0.70		-0.60	0.58	0.59	0.30	99	4	0.60		-0.27	0.43	0.24	0.49	93
5	0.60		-0.34	0.63	0.51	0.47	89	5	0.53		-0.12	0.57	0.03	0.53	93
6	0.62		-0.50	0.59	0.44	0.54	79	6	0.50		-0.50	0.10	0.49	0.51	74
7	0.24		-0.39	0.15	0.07	0.25	35	7	0.42		-0.18	0.47	0.13	0.52	49
8	0.66		-0.63	0.43	0.36	0.43	35	8	0.26		-0.39	0.25	0.29	0.52	34

WR = Word Recognition
WA = Word Analysis
OR = Oral Reading
SRC = Silent Reading Comprehension
SP = Spelling
WM = Word Meaning

* Negative correlations reflect the scoring of the *DAR* Oral Reading test. Reading miscues (number incorrect rather than number correct) are entered for the total raw score.

** Correlations at grade 1 are with the *GMRT* Level PRE Letters and Letter-Sound Correspondences subtest. Level PRE tests pre-reading concepts rather than reading itself, and so the correlations at this grade level are not as high as they are at the other grade levels, where both the *DAR* and the *GMRT* are measuring reading-related tasks.

BEST COPY AVAILABLE

Table 5

Means, Standard Deviations, and N's for the Vocabulary Test of the *Gates-MacGinitie Reading Tests* (*GMRT*) for the *DARTTS* Tryout and Validation Studies

		Mean Raw Score	SD	N
Grade 1*	Form X Tryout Sample	27.8	1.84	141
	Form Y Tryout Sample	27.7	1.66	87
	GMRT National Standardization Sample	26.9	3.06	
		Mean ESS	**SD**	**N**
Grade 1**	Validation Sample	290.62	50.38	148
	GMRT National Standardization Sample	304	61.40	
Grade 2	Form X Tryout Sample	425.50	32.50	135
	Form Y Tryout Sample	427.14	31.72	152
	Validation Sample	379.93	38.96	179
	GMRT National Standardization Sample	417	40.70	
Grade 3	Form X Tryout Sample	458.96	32.11	115
	Form Y Tryout Sample	468.62	27.23	104
	Validation Sample	431.19	35.05	177
	GMRT National Standardization Sample	458	38.50	
Grade 4	Form X Tryout Sample	487.84	37.71	99
	Form Y Tryout Sample	499.60	32.55	93
	Validation Sample	464.01	40.12	152
	GMRT National Standardization Sample	486	39.70	
Grade 5	Form X Tryout Sample	515.58	33.99	89
	Form Y Tryout Sample	515.56	35.29	93
	Validation Sample	485.32	42.98	161
	GMRT National Standardization Sample	513	42.80	
Grade 6	Form X Tryout Sample	535.14	35.57	79
	Form Y Tryout Sample	537.70	36.71	74
	Validation Sample	489.10	41.42	120
	GMRT National Standardization Sample	536	44.90	
Grade 7	Form X Tryout Sample	536.94	30.62	35
	Form Y Tryout Sample	544.78	40.03	49
	Validation Sample	526.09	56.00	86
	GMRT National Standardization Sample	557	47.30	
Grade 8	Form X Tryout Sample	567.57	34.06	35
	Form Y Tryout Sample	554.29	43.17	34
	Validation Sample	541.89	60.76	75
	GMRT National Standardization Sample	577	51.00	

* Letters and Letter-Sound Correspondences Subtest of the *GMRT* Level PRE
** *GMRT* Level R Total Score

Part II: Description of the Validation Procedures

The 1990–91 Validation Study

In the 1990–91 school year, teachers of grades 1 through 12 were asked to participate in the national validation of the complete *DARTTS* program. The teachers were asked to identify students for the study who were not achieving up to their estimated potential in reading. A total of 1216 students participated in the Validation Study. The numbers of students by grade who participated are shown in Table 6.

Table 6

Numbers of Students in the *DARTTS* Validation Study

Grade	No. of Students
1	159
2	185
3	183
4	155
5	163
6	121
7	82
8	71
9	35
10	32
11	20
12	10
Total	1216

Each teacher administered the *DAR* to individual students according to the directions provided. They also used the appropriate *TTS* lessons with the students based on the students' *DAR* results.

As a research anchor, the *Gates-MacGinitie Reading Tests* (*GMRT*) were administered to each student three times during the study. The first *GMRT* testing was in October–November 1990 just prior to the administration of the *DARTTS*. The second testing occurred in February 1991 and the third, April 1991.

The means, standard deviations, and numbers of students for the ESS scores of the first administration of the *GMRT* are provided in Table 7. In brackets underneath the average scores for the *DARTTS* Validation Study participants are the average ESSs for the participants in the *GMRT* national standardization study. The mean *GMRT* ESSs for the *DARTTS* study participants are consistently lower than those for the participants in the *GMRT* national standardization. This information illustrates that the students in the sample for the *DARTTS* Validation Study were reading about one to three years below grade level. At the upper levels, the average score differences were greater than they were at the lower levels. It may be that at the lower levels students are targeted for remediation earlier than they are at the upper grades. For example, reading two grades below grade level in high school (e.g., reading at grade 8 in tenth grade) is generally not considered to be as serious a problem as reading two grades below grade level in the primary and intermediate grades (e.g., reading at grade 2 in fourth grade).

Because there is very little difference in reading ability between average ninth and tenth grade students or average eleventh and twelfth grade students, the *DAR* tests are designed with only two levels for the high school grades, Level 9/10 and Level 11/12. Therefore, for all of the Validation Study analyses, student data for grades 9 and 10 and for grades 11 and 12 were combined.

The fall *GMRT* testing was also used to provide validity data for the *DAR* tests. The correlations between the highest mastery level on the *DAR* tests and the *GMRT* ESS scores are provided in Table 8. The table indicates a general agreement between the *DAR* and the *GMRT*.

Of particular interest are the lower correlations between the *DAR* Word Meaning test and the *GMRT* in grades 1 and 2. This is to be expected since the *DAR* Word Meaning test does not require reading, while the *GMRT* Vocabulary and Comprehension tests do. When most students acquire sufficient reading skills, at about grade 3, the correlations between the *DAR* Word Meaning test and the *GMRT* rise.

Table 9 provides the intercorrelations for the *DAR* Word Recognition, Oral Reading, Silent Reading Comprehension, Spelling, and Word Meaning tests. The Word Analysis test comprises subtests that evaluate particular aspects of the alphabetic principle of language rather than levels of a specific skill. In addition, this test is administered only at the lowest test levels. Therefore, intercorrelations between this test and the other *DAR* tests were not computed. Note the higher correlations among the tests that require reading (Word Recognition, Oral Reading, Silent Reading Comprehension, and Spelling) at all grades. Note also that the correlations of the Word Meaning test (not based on reading) and the tests that require reading are low in grades 1 and 2.

The data in Tables 8 and 9 are all based on the students who participated in the validation study. That is, these data must be interpreted in light of the overall lower reading level of the student population.

Table 7

Means, Standard Deviations, and N's for the Fall Administration of the *Gates-MacGinitie Reading Tests* —ESS Scores *GMRT* National Standardization Sample in Brackets *DARTTS* Validation Study

Grade	Vocabulary			Comprehension			Total		
	Mean	SD	N	Mean	SD	N	Mean	SD	N
1							290.62 [304	50.38 61.4]	148
2	379.93 [417	38.96 40.7]	179	347.84 [409	66.07 66.4]	178	370.52 [415	42.87 46.2]	180
3	431.19 [458	35.05 38.5]	177	414.52 [459	66.41 63.4]	176	425.54 [459	44.01 44.9]	180
4	464.01 [486	40.12 39.7]	152	447.94 [490	57.74 61.1]	151	460.08 [488	44.74 44.8]	151
5	485.32 [513	42.98 42.8]	161	476.02 [516	57.99 60.7]	161	481.09 [514	47.33 46.8]	161
6	489.10 [536	41.42 44.9]	120	473.00 [543	51.69 65.2]	120	481.90 [538	43.25 49.7]	120
7	526.09 [557	56.00 47.3]	86	508.14 [555	68.58 64.0]	87	516.93 [556	58.75 51.5]	86
8	541.89 [577	60.76 51.0]	75	529.04 [574	77.74 65.3]	75	534.48 [575	66.14 54.0]	75
9/10 9 10	581.91 [593 [609	47.29 50.9] 54.3]	67	559.93 [587 [602	71.65 65.1] 74.2]	67	572.15 [589 [605	55.63 53.9] 61.9]	67
11/12 11 12	586.90 [626 [639	37.23 55.3] 57.1]	30	544.93 [615 [629	45.66 77.4] 78.6]	30	565.80 [621 [634	38.29 64.6] 66.4]	30

Table 8

Correlations between the Highest Mastery Level on the *DAR* Tests and the Research Anchor Test (*Gates-MacGinitie Reading Tests* ESS Scores) by Grade and Test

Grade	ESS Score	Word Recognition		Oral Reading		Silent Reading Comprehension		Spelling		Word Meaning	
		r	N	r	N	r	N	r	N	r	N
1	Vocabulary										
	Comprehension										
	Total	0.63	147	0.57	138			0.51	139	0.29	125
2	Vocabulary	0.49	179	0.52	171	0.60	31	0.15	176	0.14	156
	Comprehension	0.47	178	0.45	170	0.46	31	0.15	175	-0.04	155
	Total	0.53	180	0.53	172	0.56	31	0.16	177	0.08	157
3	Vocabulary	0.69	176	0.58	170	0.58	106	0.43	166	0.49	153
	Comprehension	0.59	175	0.59	169	0.53	106	0.42	165	0.44	152
	Total	0.70	179	0.64	173	0.59	107	0.49	169	0.54	156
4	Vocabulary	0.79	151	0.68	150	0.49	124	0.46	151	0.49	140
	Comprehension	0.65	150	0.54	149	0.62	123	0.32	150	0.38	139
	Total	0.79	150	0.68	149	0.60	123	0.44	150	0.48	139
5	Vocabulary	0.70	160	0.62	156	0.66	147	0.57	154	0.63	143
	Comprehension	0.58	160	0.57	156	0.56	147	0.47	154	0.51	143
	Total	0.68	160	0.64	156	0.64	147	0.56	154	0.60	143
6	Vocabulary	0.55	119	0.56	117	0.40	111	0.49	113	0.48	109
	Comprehension	0.50	119	0.49	117	0.48	111	0.46	113	0.40	109
	Total	0.57	119	0.56	117	0.47	111	0.52	113	0.48	109
7	Vocabulary	0.72	79	0.53	79	0.68	80	0.36	79	0.63	78
	Comprehension	0.58	80	0.41	80	0.58	81	0.39	80	0.52	79
	Total	0.70	79	0.51	79	0.68	80	0.41	79	0.62	78
8	Vocabulary	0.70	70	0.52	70	0.67	70	0.58	70	0.70	68
	Comprehension	0.54	70	0.44	70	0.54	70	0.37	70	0.53	68
	Total	0.67	70	0.53	70	0.64	70	0.52	70	0.65	68
9/10	Vocabulary	0.61	67	0.56	66	0.73	66	0.54	67	0.69	65
	Comprehension	0.48	67	0.48	66	0.65	66	0.49	67	0.60	65
	Total	0.58	67	0.56	66	0.74	66	0.55	67	0.69	65
11/12	Vocabulary	0.41	29	0.38	30	0.52	30	0.19	29	0.32	30
	Comprehension	0.37	29	0.32	30	0.30	30	0.12	29	0.06	30
	Total	0.51	29	0.43	30	0.52	30	0.24	29	0.25	30

Table 9

DAR Intercorrelations of Highest Mastery Levels by Grade and Test

Grade		Word Recognition				Oral Reading			Spelling		Word Meaning
		OR	SRC	SP	WM	SRC	SP	WM	SRC	WM	SRC
1	r	0.70	—	0.69	0.24	—	0.54	0.21	—	0.22	—
	N	147	—	147	132	—	137	127	—	132	—
2	r	0.75	0.66	0.56	0.23	0.45	0.50	0.24	0.59	0.20	0.43
	N	177	32	182	162	32	174	155	32	162	29
3	r	0.75	0.72	0.66	0.49	0.55	0.57	0.40	0.42	0.33	0.53
	N	176	109	172	159	107	169	156	108	158	106
4	r	0.78	0.48	0.61	0.53	0.42	0.54	0.40	0.24	0.28	0.30
	N	153	126	153	142	125	152	142	126	142	123
5	r	0.74	0.55	0.63	0.46	0.51	0.68	0.37	0.43	0.40	0.47
	N	158	148	155	144	145	152	141	148	143	137
6	r	0.75	0.41	0.55	0.27	0.51	0.65	0.25	0.26	0.14	0.27
	N	118	112	114	110	110	113	108	111	109	107
7	r	0.72	0.62	0.60	0.51	0.52	0.51	0.49	0.35	0.38	0.67
	N	80	81	80	79	81	80	79	81	79	80
8	r	0.80	0.68	0.66	0.61	0.56	0.66	0.49	0.51	0.47	0.64
	N	71	71	71	69	71	71	69	71	69	69
9/10	r	0.80	0.64	0.61	0.51	0.63	0.64	0.54	0.63	0.50	0.49
	N	66	66	67	65	65	66	65	66	65	65
11/12	r	0.55	0.48	0.75	0.11	0.52	0.52	0.06	0.17	-0.03	0.25
	N	29	29	28	29	30	29	30	29	29	30

OR = Oral Reading
SRC = Silent Reading Comprehension
SP = Spelling
WM = Word Meaning

Part III: Featured Illustration

For this paper, to illustrate the editorial and statistical procedures that were used to create first the Validation Study edition and then the final *DARTTS* program, we will feature Level 4 of the *DAR* Word Recognition test. Important factors influenced the decision to feature this particular test at this particular level.

The Word Recognition test plays a dual role in the program. This test provides valuable information about the student's reading vocabulary, but it also acts as a "routing test" into the four remaining *DAR* tests that also evaluate reading and language: Word Analysis, Oral Reading, Silent Reading Comprehension, and Spelling. Only the Word Meaning test, because it is not a test of reading, is started based on the student's school grade placement rather than functioning reading level.

Level 4 represents an important transition in reading development. It is the beginning of what Chall has called "Stage 3: Reading for Learning the New" (Chall, 1983). From the beginning of reading instruction through approximately grade 2 or 3, while children are extracting meaning from written text, they are also learning the alphabetic principle of language, and so decoding skills play an important role in their reading comprehension. Somewhere around grade 3 or 4, the content and language of the materials students are expected to read becomes less familiar, and so extracting new information and analyzing that information becomes increasingly important. In the *DARTTS* program, if students achieve mastery on the Level 4 Word Recognition test, it usually means they have mastered the concepts that are tested in the *DAR* Word Analysis test, and so, at the examiner's discretion, they may by-pass this test and go directly to Oral Reading. (See Table 1 on page 2.)

Item Selection for the National Item Tryout Study

The items (words and reading passages) selected for the *DAR* were compared for difficulty and grade placement based on various standard sources—word lists, readability measures, and grade placement within widely used curriculum materials. This item selection discussion will focus on the Word Recognition test, particularly on Level 4 of this test.

The words for the *DAR* Word Recognition test were compared with standard word lists such as the Thorndike-Lorge frequency word count (Thorndike & Lorge, 1944), the Harris-Jacobson list based on frequency of use in basal readers (Harris & Jacobson, 1982), and on the Ginn combined grade level based on oral use, textbooks, children's readers, etc. (Johnson, Moe, & Baumann, 1983). In addition, the words were checked for their inclusion on the Spache list of 1000 common words, which is a list known by first and second graders (Spache, 1974), and the Dale List of 3000, which is a list known to 80 percent of fourth graders (Dale & Chall, 1948). For Levels 3 and above, the difficulty of the words was also recorded from the Dale-O'Rourke *Living Word Vocabulary* (Dale & O'Rourke, 1981). Table 10 provides the source information for all of the words on the final Level 4 Word Recognition test.

These sources served only as general guidelines. In many cases, as the table shows, the sources do not agree about the difficulty of words. For example, *celebrate* is on the Dale list, but rated as a grade 5–6 word by Thorndike-Lorge and a grade 3 word by Harris-Jacobson.

In cases where the *DARTTS* authors' knowledge of students' abilities did not agree with the source information, the authors' determination as to the proper placement of the word in the *DAR* prevailed. It is interesting to note that in

Table 10

Source Information for the Level 4 *DAR* Word Recognition Test

Level IV	Thorndike-Lorge	Harris-Jacobson	Spache 1000	Dale 3000	Ginn Overall	Dale-O'Rourke
straight	1-2	2	yes	yes	2	4(89%)
celebrate	5-6	3	no	yes	4	4(86%)
favorable	4	5	no	no	4	8(77%)
exhibit	4	4	no	no	4	4(72%)
creation	4	4	no	no	4	6(93%)
inspiration	5-6	5	no	no	6	6(79%)
survival	7-8	5	no	no	5	6(77%)
inhabitant	4	6	no	no	5	8(76%)
tragedy	4	4	no	no	4	6(88%)
permanent	4	4	no	no	4	6(78%)

most cases our research confirmed the authors' expert judgments. For example, *survival* does not appear to be a Level 4 word based on the sources, but our data, both from the Tryout Study and from the Validation Study, confirmed the authors' difficulty estimation. The grade 4 *p*-value for *survival* was 58 for the Tryout Study and 54 for the Validation Study, confirming the authors' determination that this word is well with the grasp of average fourth grade students. (See Tables 11 and 12.)

For the Tryout Study, eight words were selected for each level of Form X and Form Y.

Item Selection for the Validation Study

Based on the data from the Tryout Study and also once again the authors' expert determinations, words were selected for the Validation Edition test levels. From the Tryout Study pool of words, words that demonstrated the most growth potential across several levels were preferred. For example, the word *exhibit* was quite difficult for grade 3 students, a bit easier for grade 4 students, and quite easy for grade 5 students, so it was selected for the Level 4 Word Recognition test. (See Table 11.)

In general, words that 100% of the on-level students could read correctly were not selected for the Word Recognition test. The exception to this standard was for the lowest levels of the test, which contain mostly easy sight words, such as *see*, *go*, *car*, and *ball*.

The words were positioned in the Validation Study edition according to their relative difficulties. The following criteria governed the arrangement of words by and within level:

- Within each level, the words should progress from easy to difficult.
- Across all of the levels, the words should progress in the general direction of easy to difficult.
- Each higher level should begin with a word that is slightly easier than the most difficult word on the previous level.

The last criterion was important. The Word Recognition test is the student's first experience with the *DAR*. While the general direction of words across all of the levels is from easy to difficult, because students enter the test at all levels, it is important to begin testing with a relatively easy word. This initial success experience facilitates the testing process as students begin to discover that there are things that they know. As an example, the most difficult word in the Level 4 test is *permanent*, which had a Tryout Study *p*-value of 47 at grade 4. The first word in the Level 5 test is *ancestor*, which had a Tryout Study *p*-value of 68 at grade 4 and 81 at grade 5, making it slightly easier than *permanent* based on our research. In a similar fashion, the last word on the Level 5 Word Recognition test is slightly more difficult than the first word on the Level 6 test. In addition to making it easier for teachers and students to enter the test at any point, this method of ordering words also provides students and teachers with a brief relaxation as they progress from level to level.

Table 11 shows the Tryout Study item difficulties by grade level for the Level 4 *DAR* Word Recognition test words. Of the ten words that were selected for the final Level 4 test, nine were tried out in grades 3, 4, and 5, and one (*inspiration*) was tried out in grades 4, 5, and 6. The words appear in the table as they were ordered for the Validation Study, based on the Tryout Study data. As stated before, data were used to guide rather than to make decisions. Both the data and the authors' expert judgments were critical to the success of the decision making.

The data analysis for the Validation Study was necessarily more complex than that done for the Tryout Study. It was essential for the acceptance of the program that reliable information regarding the relative difficulties of all items be provided. The students' school grade placement was less important than their actual performance for this purpose. To get a true picture of students' current achievement, it would be important that items be ordered in the test by difficulty, with easier levels preceding more difficult levels, but also with easier items preceding more difficult items within a given level.

A complicating factor was our student sample for the Validation Study, which consisted of students who were reading, on average, one to three years below their grade placement level (see Table 5 on page 8). On-grade-level data would probably not provide the information required to validate the program. However, because the target population for the program would be primarily students who are experiencing difficulty learning to read, it was important to validate the program on that population.

To analyze how items were functioning across test levels and also within each level, the distributions of item responses were analyzed. These distributions comprised the percent of students who answered each item correctly for that level (*p*-values). It was assumed that as items got more difficult, fewer students would answer the item correctly. Then different analyses were performed:

- Initially the responses of students by grade level were examined. However, a simple visual examination of the students' *Response Record* booklets for the Validation Study revealed that there were students at all grade levels who were achieving mastery on several test levels before they were reaching their highest mastery level. We did not want the ease of some students' achievement to contaminate the data for any item. Further analysis would be needed.

- It was at this point that we decided to break down the data based on whether or not students mastered a level. We were most interested in finding out exactly where in the entire scope of the test that students began to fail. The mastery/nonmastery information gave us that critical information.

Table 12 shows how the Validation Study was able to refine the Tryout Study data through the analysis of students' performance based on mastery or nonmastery of the test level. The words are the same as those shown in Table 11, however, here the words are listed in the order in which they were published in the final test form. The "Mastery" column under each grade lists the *p*-values of the words for students who mastered this level and went on to the Level 5 test. The "Nonmastery" column shows the same information for those students who did not master this level, and the "Total" column includes all of the students in the sample, regardless of whether or not they mastered the level.

The numbers in parentheses after each *p*-value show the order by difficulty (based on the *p*-value) of the words for each column. In general there is a great deal of agreement in the order of the words among the different columns and across the grades. However, exceptions are obvious. For example, in grades 3, 4, and 5, all of the students in the "Mastery" column were able to read the word *straight*, placing this word at number 1 (least difficult) in all of these grades. However, in grade 6 *straight* is placed at number 2 because, even though it is a fairly easy word, one or more grade 6 students in the sample who mastered this level read *straight* incorrectly.

Another situation where the order by difficulty for a word makes it stand out in the column is for a word such as *exhibit*. In the "Total" column, *exhibit* is positioned at either "4" or "5" (moderately easy) until grade 6, where its position becomes "9" (quite difficult). This word is an example of a word that pointed us in the direction of separating the mastery from the nonmastery data. As stated above, we needed to know exactly where students *began* to fail. For the word *exhibit*, the Nonmastery order by difficulty is between "3" and "5" all the way across the grades, indicating a more realistic placement range for this word for all students. The authors chose to place this word in the number 4 position in the final test form.

A close reading of Table 12 reveals other differences in word ordering by difficulty across the grades. The source information provided in Table 10 shows that even highly respected standard sources can disagree regarding the difficulties of words. Our research showed that some words appeared to be easier for less accomplished than for more accomplished students. For example, in Table 12, examining the Nonmastery columns shows that the word *survival* appears to be the most difficult of these words at grade 3, but much easier at the higher grades. The reverse is true of the word *permanent*, which looks very difficult at grades 4–6 and considerably easier at grade 3. In situations such as these, we chose to look primarily at the target grade level for the test level, in this case grade 4 for Level 4. In addition, we also relied to a great extent on our authors' expert judgments.

Table 11

DAR Tryout Study Level 4 Word Recognition Test Item Difficulties

Grade	3	4	5	6
straight*	79	93	98	
celebrate*	76	87	97	
exhibit	59	81	84	
favorable*	54	80	91	
inspiration*		69	80	82
creation	47	68	77	
tragedy	47	65	70	
survival	28	58	81	
inhabitant	43	54	80	
permanent	39	47	62	
N Form X*	134	110	108	
N Form Y	122	106	108	77

Table 12

DARTTS Validation Study Level 4 Word Recognition Test Item Difficulties

Grade	3			4			5			6		
	Mastery	Nonmastery	Total	Mastery	Nonmastery	Total	Mastery	Nonmastery	Total	Mastery	Nonmastery	Total
straight	100 (1)	89 (1)	92 (1)	100 (1)	97 (1)	98 (1)	100 (1)	87 (1)	93 (1)	98 (2)	79 (2)	91 (2)
celebrate	94 (3)	79 (2)	83 (2)	98 (2)	84 (2)	91 (2)	100 (1)	87 (1)	93 (1)	100 (1)	92 (1)	97 (1)
favorable	100 (1)	36 (5)	54 (4)	94 (3)	57 (3)	74 (3)	96 (3)	52 (3)	74 (3)	91 (4)	63 (3)	81 (3)
exhibit	89 (5)	40 (3)	54 (4)	92 (4)	41 (4)	65 (4)	91 (4)	39 (5)	64 (5)	74 (9)	38 (4)	61 (9)
creation	83 (6)	23 (8)	40 (6)	90 (6)	34 (5)	61 (5)	89 (5)	35 (6)	62 (6)	95 (3)	29 (7)	72 (4)
inspiration	94 (3)	40 (3)	55 (3)	78 (7)	33 (6)	54 (6)	75 (8)	24 (8)	50 (8)	88 (5)	29 (7)	67 (6)
survival	78 (8)	13 (10)	31 (10)	92 (4)	21 (8)	54 (6)	85 (6)	28 (7)	56 (7)	86 (7)	38 (4)	69 (5)
inhabitant	72 (9)	26 (6)	38 (7)	75 (9)	26 (7)	49 (8)	85 (6)	50 (4)	67 (4)	81 (8)	38 (4)	66 (7)
tragedy	83 (6)	21 (9)	38 (7)	78 (7)	17 (9)	46 (9)	68 (9)	9 (10)	38 (9)	88 (5)	21 (9)	64 (8)
permanent	72 (9)	26 (6)	38 (7)	65 (10)	10 (10)	36 (10)	57 (10)	15 (9)	36 (10)	65 (10)	13 (10)	46 (10)
N	18	47	65	51	58	109	53	54	107	43	24	67

Educational Importance of the Study

Teachers require a variety of tests that will assess accurately and completely what their students know. At the same time, parents, school administrators, and other education professionals rightly insist that the tests they use be proven to be reliable and valid. Test publishers have the responsibility of fulfilling the expectations of all of these groups.

The educational community is currently quite interested in performance assessment, in tests that will allow students to demonstrate what they know. Educators know that they can rely on the accuracy of the normative data and the reliability and validity information publishers provide with standardized multiple-choice tests. If students' scores on performance-based instruments are to be equally trusted, these newer test types will have to provide equally valid and reliable information.

This paper describes how we analyzed mastery and nonmastery data to validate the *DARTTS* program. However, the methodology we used may have broader applications to research on other such instruments. The procedures were effective for our purposes; they are also simple, which should add to their attractiveness and applicability for the future.

References

Chall, J. S. (1983). *Stages of reading development* (pp. 20-23). New York: McGraw-Hill Book Company.

Chall, J. S., Jacobs, V. A., & Baldwin, L. E. (1990). *The reading crisis: Why poor children fall behind*. Cambridge, MA: Harvard University Press.

Dale, E., & Chall J. S. (1948). *A formula for predicting readability*. Columbus, OH: Bureau of Educational Research, Ohio State University.

Dale, E., & O'Rourke, J. (1981). *The living word vocabulary*. Chicago: World Book-Childcraft International.

Harris, A. J., & Jacobson, M. D. (1982). *Basic reading vocabularies*. New York: Macmillan Publishing Company.

Johnson, D. D., Moe, A. J., & Baumann, J. F. (1983). *The Ginn word book for teachers*. Columbus, OH: Ginn and Company.

MacGinitie, W. H., & MacGinitie, R. K. (1989). *Gates-MacGinitie reading tests technical report*. (Table 25, p. 45). Chicago: The Riverside Publishing Company.

Roswell, F. G., & Chall, J. S. (in press). *DARTTS technical manual*. Chicago: The Riverside Publishing Company.

Spache, G. D. (1974). *Good reading for poor readers*. Champaign, IL: Garrard Press.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College Press.